ISSN 0840-8440

PROCEEDINGS

TECHNOLOGY TRANSFER CONFERENCE 1988

November 28 and 29, 1988

Royal York Hotel

Toronto, Ontario

SESSION D

ANALYTICAL METHODS

Sponsored by

Research and Technology Branch

Environment Ontario

Ontario, Canada

AASG

Copyright Provisions and Restrictions on Copying:

This Ontario Ministry of the Environment work is protected by Crown copyright (unless otherwise indicated), which is held by the Queen's Printer for Ontario. It may be reproduced for non-commercial purposes if credit is given and Crown copyright is acknowledged.

It may not be reproduced, in all or in part, for any commercial purpose except under a licence from the Queen's Printer for Ontario.

For information on reproducing Government of Ontario works, please contact ServiceOntario Publications at copyright@ontario.ca

DP11

ROBUSTNESS OF SIMPLE HYPOTHESIS TESTING METHODS WITH CENSORED ENVIRONMENTAL QUALITY DATA

E.E. Creese

Creese Environmental Consulting,
P.0. Box 91, Waterloo, Ontario
N2J 3Z6

INTRODUCTION

Frequently when a chemical parameters of environmental concern (e.g. chlorinated organics or heavy metals) is measured, a sample will contain some observations below the analytical detection limit. Such data is termed a *censored* data set.

Previous studies (Gilliom & Helsel, 1986; Helsel & Gilliom, 1986) have shown that, of the methods that they investigated, log-probability regression to the curve of a normal distribution was the most robust for estimating population parameters such as the mean and standard deviation of parent populations of environmental data. This work takes as a starting point the assumption that the lognormal distribution is the most likely parent distribution of chemical water quality data. It endeavors to assess the reliability of simple hypothesis testing, such as the z test or the t test, when population or sample parameters are estimated by log-probability regression.

The basic procedure is to generate a large number of simulated samples, all with a specified number of observations. Then censoring is performed at a certain defined cut-off point. The mean and variance of the population (or sample, depending on the method) are estimated by log-probability regression. Then, a hypothesis test is performed using the estimated mean and variance to determine if the sample mean can be inferred to be equal to the population mean. Since the population mean is known, which, of course, is the whole point of using simulated data, one can say whether or not the statistical test failed. Such a failure, that of rejecting a true null hypothesis, is termed a type I error. This work involves the use of Monte Carlo methods to compare to compare nominal type I error rates to actual type I error rates. At present, several regression methods are being compared, but only one is reported on here. The work is still in progress. For this reason, methods are the main topic of discussion in this paper. Only preliminary results are available.

METHODS

The work is being carried out on a Macintosh Plus microcomputer with a 20 megabyte hard drive. Random number generation and statistical routines are performed with the Systat application. All the Systat calculations are done in double precision, an automatic feature of most Systat routines. Iterative Systat programs are assembled by a Hypercard program and submitted for batch execution to Systat using the Macromaker utility. Macromaker returns control to Hypercard, which then composes the next set of Systat programs, and so on.

Generation of random numbers

The algorithm used by Systat for the generation of random numbers is given in Wichman & Hill (1982). It has a cycle length of 2.78×10^{13}. On opening the Systat DATA module, the random number generator always starts at the same point in the cycle. It is possible to provide a seed number between 1 and 30,000 that can be entered to make the generator start at a different point. This, in effect, reduces the cycle to 30,000, supposing, of course, that the choice of the seed is random. For this reason, a batch of 10,000 random numbers with the required probability distribution was produced initially in a single session in the DATA

module. When, say, another 10,000 are required, the Systat DATA module will be opened, 20,000 random numbers will be produced, discarding the first 10,000.

Generation of random samples

As mentioned above, a set of 10,000 lognormally distributed random numbers, x_i, was produced. The population median was set to $v_x = 1$ and the coefficient of variation to $cv_x = 1$. The first step in generating the numbers was to generate the corresponding normally distributed numbers, y_i, where $x_i = \exp[y_i]$. Normal random samples were simulated using the equation:

$$y_i = \mu_y + \sigma_y \varepsilon_i$$

In this equation, ε is the error term, which is a normally distributed random variate with a mean of 0 and a variance of 1. It was provided by the Systat random number generator. It is apparant that:

$$\mu_y = v_y = \ln[v_x] = 0.$$

The relationship between σ_y and cv_x is given by Aitchison & Brown (1957):

$$cv_x^2 = \exp[\sigma_y^2] - 1.$$

Samples of sample size n = 10 were abstracted from the set thus generated, the first 10 numbers becoming the first sample, the next 10 becoming the second sample, etc.

Censoring

Four censoring points were chosen to correspond to the 20th, 40th, 60th and 80th percentiles of the parent population. These correspond to four hypothetical detection limits, x_D, of chemical analysis. They are calculated from the normal distribution as follows:

$$x_D = \exp[\mu_y + z[p]\ \sigma_y] \mid p \in \{.2,.4,.6,.8\}$$

Table 1. Detection limits, x_D, corresponding to the 20th, 40th, 60th and 80th percentiles of a lognormally distributed quantity with median value and coefficent of variation both equal to 1.

Percentile	x_D
20	0.496
40	0.810
60	1.235
80	2.015

Baseline Check

The set of 10,000 random numbers generated were tested for compliance with the original population parameters that defined it. This was done to ensure that the computer programs were indeed written and executed according to plan, in other words, to make sure that everything was done right. To accomplish this, 1,000 samples of n = 10 were tested in turn by a Student's t test in Systat module, STATS, against the true hypothesis that $\mu_y = 0$. Unlike SAS, or SPSS, which determine the probabilies corresponding to t by interpolating from a table, Systat calculates the probability, resulting in greater accuracy. The algorithm for this calculation is decribed in Lund & Lund (1983).

The results of this check are shown in Table 2. At a nominal type I error rate of $\alpha = 5\%$, it is to be expected that the t test will fail 5% of the time. Table 2 shows that the actual type I error was a = 4.9%, which is well within the 99.9% confidence limits of the nominal type I error rate. The 99.9% confidence interval of α was determined from the following expression:

$$\alpha \pm z[.9995]\ (\alpha\ (1\text{-}\alpha)\ /\ m)^{0.5}$$

where m is the number of tests, in this case, 1,000.

Table 2. Results of 1,000 t tests performed on simulated environmental quality data. α is the nominal type I error rate and a is the actual type I error rate.

α	a	99.9% confidence interval of α
5%	4.9%	4.8% - 5.2%

Simulation Runs

In these runs, simulated samples of n = 10 were ordered by rank. Then they were censored at one of the four points listed above in Table 1. Normal scores, z, were then computed for each of the remaining sample observations by:

$$z_i = z[R_i\ /\ (n + 1)],$$

where R_i is the rank of the observation. Regression of ln[x] on z was then done according to the model,

$$\ln[x] = m_y + s_y\ z,$$

to obtain estimates of the population mean, μ_y, and standard deviation, σ_y. At higher levels of censoring, some samples had to be discarded, since regression could not be performed if less than two sample observations remained above x_D. After regression, hypothesis testing was done, choosing as the null hypothesis that the sample mean is equal to the population mean. Type I errors were counted, and at the end of each run, a, the actual type I error rate was calculated.

RESULTS

To date, very few data generating runs have been completed. So far hypothesis testing has been limited to a z test. A confidence interval for the appropriate probability, p, is constructed about the estimate of population mean, thus:

$$m_y \pm z[p]\ s_y\ /\ \sqrt{n} \mid p = 1 - \alpha\ /\ 2$$

A type I error is indicated if the true population mean, μ_y, falls outside the confidence interval.

Table 3. Results of z tests on 500 samples of sample number n = 10, performed on simulated environmental quality data. The nominal type I error rate is $\alpha = 5\%$. a is the actual type I error rate, m is the number of z tests actually performed.

censoring level	a	m	95% confidence interval of α
0%	4.8%	500	4.0% - 6.0%
20%	6.2%	500	4.0% - 6.0%
40%	12.8%	499	4.0% - 6.0%
60%	26.7%	475	4.0% - 6.0%
80%	50.5%	325	3.8% - 6.2%

The trend seen in Table 3 is of course what would be expected. The power of the z test to distinguish the correct hypothesis decreases with the level of censoring. It can be seen that at 80% censoring, the z test has lost all discriminating power. It is basically a 50:50 chance whether or not it picks the correct hypothesis.

DISCUSSION

More preliminary experimentation is required before production of tables can begin. It will be necessary to know, for example, whether the absolute number as well as the proportion of observations above detection influences the actual type I error rate.

The ultimate goal of this work is to generate tables so that one could perform a standard hypothesis test at a given nominal type I error rate and then look up what the actual type I error is. Alternatively, one could decide, given the degree of censoring, what nominal type I error to test a hypothesis at in order to obtain the desired actual type I error probability. Such tables would be of great use to practising environmental scientists.

REFERENCES

Aitchison, J. & J.A.C. Brown. 1957. *The Lognormal Distribution with special reference to its uses in economics*. Cambridge University Press.

Gilliom, R.J. & D.R. Helsel. 1986. Estimation of distributional parameters for censored trace level water quality data. I. Estimation techniques. *Water Resources Research, 22(2):* 135-146.

Helsel, D.R. & R.J. Gilliom. 1986. Estimation of distributional parameters for censored trace level water quality data. 2. Verification and applications. *Water Resources Research, 22(2):* 147-155.

Lund, R.E. & J.R. Lund. 1983. Probabilities and upper quantiles for the studentized range. Algorithm AS 190. *Applied Statistics, 312:* 204-210.

Wichman, B.A. & I.D. Hill. 1982. An efficient and portable pseudo-random number generator. Algorithm AS 183. *Applied Statistics, 311:* 188-190.


(8177)
TD/5/T43